# SEARCHING DOCUMENTS USING A DIMENSIONAL DATABASE

## CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of provisional U.S. Application No. 60/237,672, filed October 3, 2000, entitled, "MULTIDIMENSIONAL SEARCH OF DOCUMENTS" which is hereby incorporated by reference.

## TECHNICAL FIELD

[0002] The present invention is directed toward the field of document searching, and, more particularly to the field of web-based document searching.

## BACKGROUND OF THE INVENTION

[0003] The World Wide Web ("the web") is a distributed international electronic library of documents and other data resources. A particular document is accessed on the web using a unique identifier for the document, called a "URL," short for Uniform Resource Locator. If a user seeking to access a particular document has the URL for the document, s/he may simply type it into the URL field of a web browser. In many cases, the URL for a document may be obtained from a second, related document containing a link to the first document.

[0004] It is conservatively estimated that over a billion documents are available on the web. (Indeed, smaller "webs," such as "Intranets" used only by the employees of a particular business, may themselves provide access to hundreds of thousands of documents.) For a particular user having a particular need, the web may contain several documents that address the need, all unknown to the user. For example, for a user interested in details of the 1955 grape harvest in Eastern Washington, 15 documents may be available on the web that contain such information, all unknown to the user.

[0005] In order to help users identify documents on the web relating to particular subjects, hierarchical web directories and web search engines have been developed. A hierarchical web directory is a set of human-compiled lists of documents available via the web each relating to a particular subject represented in a hierarchy of topics. Table 1 below shows a designation of a hierarchical web directory topic corresponding to a list of documents available via the web that includes documents containing information about the 1955 grape harvest in Eastern Washington.

Society and Culture

Food and Drink

Spirits

Wine

Regional

Eastern Washington

TABLE 1

[0006] The topic corresponding to the list, Eastern Washington, is a subtopic of the topic "Regional," which is a subtopic of the topic "Wine," which is a subtopic of the topic "Spirits," which is a subtopic of the topic "Food and Drink," which is a subtopic of the topic "Society and Culture." In order to provide a hierarchical web directory, its provider must create a hierarchy of topics, identify documents available via the web, and identify topics to whose lists the identified documents should be added.

[0007] A web search engine, on the other hand, allows users to type one or more key words and returns a list of documents containing those keywords. In particular, web search engines typically include documents in the list that have the highest percentages of occurrences of the key words among all of the documents. For example, to identify documents containing details of the 1955

grape harvest in Eastern Washington, a user might type the key word string "1955 grape harvest Eastern Washington." The web search engine processes such queries against a database representing the contents of as many web pages as possible, typically gathered by "spidering," or automatically traversing links from known web pages to new web pages.

[0008] Both of these conventional approaches to identifying documents on the web have significant disadvantages. Hierarchical web directories are extremely labor intensive, requiring human editors to review and categorize web documents. This reliance on manual processes often results in outdated or inaccurate content. Also, hierarchical web directories are only usable to identify web pages relating to topics created by human editors. Hierarchical web directories are also difficult for users to successfully use, as a user must typically select the exact same sequence of subtopics as the person that catalogued the web site.

[0009] Web search engines, while not typically suffering from the deficiencies of hierarchical web directories relating to their manual nature, have the disadvantage that they rely on the occurrence of particular key words in sought web pages. Because many words have multiple meanings, Web search engines often generate false positive matches, where a keyword appears in the Web page in a different sense than the sense intended by the user in formulating the query. On the other hand, because of the large number of words that can be used to get across the same idea, Web search engines also often generate false negative matches. Web search engines also typically filter out noise words that occur in most web pages, such as "an" or "if," which make it impossible to search for web pages using these words when using a web search engine. Further, aside from applying certain frequency analysis techniques, web search engines typically ignore the specific usage and significance of particular key words in the searched web pages.

[0010] Accordingly, a more effective approach to identifying documents on the web and in other electronic libraries would have significant utility.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Figure 1 is a high-level block diagram of the computer system upon which the facility preferably executes.

[0012] Figure 2 is a flow diagram showing the steps preferably performed by the facility in order to construct a dimensional model.

[0013] Figure 3 is a diagram showing a registration form completed by the webmaster of a company root page.

[0014] Figure 4 is a diagram showing the source code for a sample company root page as automatically located and parsed by the facility.

[0015] Figure 5 is the data structure diagram showing the definition of a sample dimensional model in dimensional database notation.

[0016] Figure 6 is a data structure diagram showing the data tables that preferably underlie the sample dimensional model shown in Figure 5.

[0017] Figure 7 is a flow diagram showing the steps preferably performed by the facility in order to process search requests against the dimensional model.

[0018] Figure 8 is a diagram showing a sample search request.

[0019] Figure 9 is a diagram showing a representation of the search request shown in Figure 8 expressed in a search request star notation.

[0020] Figure 10 is a diagram showing a second sample search request for the same dimensional model.

[0021] Figure 11 is a diagram showing the second sample search request expressed in search request star notation.

[0022] Figure 12 is a diagram showing a search request, expressed in star notation, for a yellow pages entry, *i.e.*, an entry in a business telephone directory.

[0023] Figure 13 is a diagram showing a search request, expressed in search request star notation, for descriptions of products sold by a computer superstore.

[0024] Figure 14 is a diagram showing a search request, expressed in search request star notation, for a library resource, such as a book or an audio or video recording.

## DETAILED DESCRIPTION

[0025] Embodiments of the present invention provide a software facility for selecting documents or other types of resources using a dimensional model ("the facility"). In some embodiments, the facility is adapted to select web pages in response to user queries.

[0026] The facility preferably generates and maintains a dimensional model of a number of documents, such as web pages. For each modeled document, the model contains information about the document in at least a portion of a number of different informational dimensions. For example, for documents that are home pages of companies, the facility preferably includes in the model for each modeled document indications of the name, type, category, and location of the company. The facility preferably obtains this information using one or more of a variety of techniques, including manual or automatic submission, or automatic spidering combined with parsing and/or natural language understanding.

[0027] The facility preferably maintains the information obtained for modeled documents in a dimensional model. The dimensional database techniques discussed herein are well-known to those skilled in the art, and are described in greater detail in Kimball, Ralph: "*The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses,*" 1996. The model includes a fact table that contains a row for each of the modeled documents. For example, the fact table may be comprised of rows each corresponding to a company home page, and containing the URL for that company home page. The model further includes a dimension table for each of the modeled dimensions. For example, the model may include four dimension tables, one each for the company_name, company_type, company_category, and company_location dimensions. Each dimension table has a row for each unique value of the dimension. For example, the company_location dimension table may have a row for each unique location of one or more modeled company root pages.

[0028] The model may be normalized by joining all of the rows of all of the dimension tables to the fact table, generating a result table in which each row explicitly contains all of the information about one document. This result table, however, can be many times larger than the entire model, and is less efficient than the model for use in executing queries.

[0029] The facility receives queries that specify functions on at least some of the modeled dimensions. For example, the facility may receive a query that specifies values for three of the four modeled dimensions. To process the query, for each specified dimension value, the facility selects the rows of the corresponding dimension table that match the value. The facility then joins the selected rows in the dimension tables to the fact table. The result table from this join is generally of manageable size, and contains a row for each document that satisfies the query, and is preferably used by the facility to generate a query result that the user can use to access the documents that satisfies the query.

[0030] In this manner, document queries that more effectively describe sought documents may be processed without requiring significant manual effort on the part of the operators of the facility or relying on the adequacy of such effort, producing more useful query results than conventional document searching approaches.

[0031] Figure 1 is a high-level block diagram of the computer system upon which the facility preferably executes. The computer system 100 contains one or more central processing units (CPUs) 110, input/output devices 120, and a computer memory (memory) 130. Among the input/output devices is a persistent storage device 121, such as a hard disk drive, and a computer-readable media drive 122, which can be used to install software products, including components of the facility, which are provided on a computer-readable medium, such as a CD-ROM. The input/output devices also include a network connection 123, through which the computer system 100 may be connected to the network to be analyzed by the facility. The memory 130 preferably contains the dimensional search facility 131, as well as dimensional models of document sets 132 and 133, both preferably

generated and used by the facility. While items 131-133 are preferably stored in memory while being used, those skilled in the art will appreciate that these items, or portions of them, may be transferred between memory and the persistent storage device for purposes of memory management and data integrity. While the facility is preferably implemented on a computer system configured as described above, those skilled in the art will recognize that it may also be implemented on computer systems having different configurations, or distributed across multiple computer systems.

[0032] Figure 2 is a flow diagram showing the steps preferably performed by the facility in order to construct a dimensional model. Those skilled in the art will appreciate that these steps could be used to construct several different dimensional models simultaneously. In step 201, the facility obtains a registration record containing the URL for a company root page or other modeled document, as well as dimension values describing the company, or other attributes of the document. The facility preferably obtains this registration information in a variety of different ways, including receiving a registration form filled out by someone responsible for the document, such as the webmaster for a company root page, or such as by automatically exploring and analyzing documents.

[0033] Figure 3 is a diagram showing a registration form completed by the webmaster of a company root page. The form 310 contains a field 320 in which the URL used to identify the company root page has been entered. It will be appreciated by those skilled in the art that the facility may use many different types of references, or "links," to refer to modeled documents and other resources. For example, URLs specified in accordance with RFC 1738 may be used to refer to online resources, as may various other electronic naming conventions. Additionally, various schemes, such as street addresses or Dewey decimal numbers, may be used to refer to offline resources. Fields 321-324 are fields into which attribute values for the company root page have been entered.

[0034] Figure 4 is a diagram showing the source code for a sample company root page as automatically located and parsed by the facility. Source code segment

400 shows the first eleven lines of the source code. As the source code was retrieved by the facility using the URL for the company root page, the URL for the source code is known to the facility. Attribute values for the company root page are specified on lines 3-6 in the XML block occurring in lines 2-7. These lines are preferably inserted into the HTML source code by the webmaster of the company root page to support indexing by the facility, as well as for a variety of other purposes. Embedding XML source code within HTML source code in this manner has been proposed by some commentators to support the inclusion of XML data in existing HTML documents. Alternatively, attribute information may be included in existing HTML constructs, such as the HTML  tag. For example, the attribute value information shown on line 3 would be encoded using the  tag as follows:

[0035] " Hughes Satellite, Inc. "

[0036] Where the company root page is expressed in XML, the attribute information is merely included in custom attribute tags that are added by the webmaster to the XML document comprising the root page.

[0037] Returning to Figure 2, in step 202, the facility adds a row to the fact table of the dimensional model for the registration record obtained in step 201. The facility further adds rows to dimension tables of the dimensional model as necessary. After step 202, the facility continues in step 201 to obtain the next registration record.

[0038] Figure 5 is the data structure diagram showing the definition of a sample dimensional model in dimensional database notation. The model has a single fact table 510, and four dimension tables 520, 530, 540, and 550. Dimension table 520 is for the company_name dimension and contains a company_name field 522 and a company_name_key field 521 in each row. Dimension table 530 is for the company_dimension, and contains a company_type field 532 and a company_type_key field in each row. Dimension table 540 is for the company_location dimension, and contains a company_location field 542 and a company_location_key field 541 in each row. Dimension table 550 is for the

company_category dimension, and contains a company_category field 552 and a company_category_key field 551 in each row. The fact table 510 contains a row for each modeled document. Each row contains a company_name_key field 511, which contains the key corresponding in the company_name dimension table 520 the value of the company_name attribute for the company root page to which the row of the fact table corresponds. The line between fields 511 and 521 indicates that tables 510 and 520 can be joined on the company_name_key field. Such a join operation, if applied to all of the rows of both tables, would produce a result table containing all of the information in the fact table, plus an additional column containing the company_name field for each row. The fact table further contains key fields 512, 513, and 514, through which the fact table may be joined to dimension tables 530, 540, and 550, respectively. Finally, fact table 510 includes a company_root_page_link field 515 that, in each row, contains a link to the company root page to which the row corresponds. In one embodiment, in which two or more documents may have all the same attribute values, a row of the fact table may correspond to more than one document. In this case, rows of the fact table that correspond to more than one modeled document preferably include the URL to a web page containing a list of these documents rather than a URL directly to a particular document. In an additional embodiment, rows that contain such a URL further include an indication of the number of documents listed at that URL. This indication may be a count of these documents, may indicate a range into which the number of documents falls, or may simply indicate whether the number of listed documents makes the list of documents too large to retrieve, at least under certain circumstances.

[0039] In some embodiments, the facility supports hierarchical dimensions. For example, rather than merely reflecting a city, a more thorough hierarchical notion of the company_location dimension as shown below in Table 2 may be used:

Country

State

City

TABLE 2

[0040] Such hierarchical dimensions may be represented either in a single dimension table containing a column for each hierarchical component of the dimension, *i.e.*, a single company_location dimension table containing separate columns for country, state, and city, or by using a "snowflaked" sequence of dimension tables emanating from the fact table in decreasing level of detail, *i.e.*, a dimension table for city, referred to by the fact table and referring to a dimension table for state, which in turn refers to a dimension table for country.

[0041] Figure 6 is a data structure diagram showing the data tables that preferably underlie the sample dimensional model shown in Figure 5. The diagram shows fact table 610, as well as dimension tables 620, 630, 640, and 650. The fact table 610 has columns 611-614, each corresponding to the key of one of the dimension tables. For example, sample row 616, corresponding to the sample company root page whose attribute values are shown in Figures 3 and 4, contains the company_name_key 237, as does row 626 of dimension table 620 for the company_name dimension. That row of dimension table 620 indicates that the company root page to which row 616 of the fact table corresponds has the value "Hughes Satellite, Inc." for the company_name attribute. The fact table 610 further includes column 615, which contains in each row the URL of the company root page to which the row corresponds, in this case, "http://www.hughessatellite.com". When the facility obtains the attribute information for this company root page, it adds row 616 to the fact table 610. It further searches each of the dimension tables to determine whether the obtained attribute value for that dimension is already contained in the dimension table. If

so, the facility merely copies the key for that row of the dimension table into the row being added to the fact table. If the attribute value is not already in the dimension table, then the facility adds a row to the dimension table containing the attribute value and a key that is unique within the dimension table, and adds that key to the row being added to the fact table.

[0042] Figure 7 is a flow diagram showing the steps preferably performed by the facility in order to process search requests against the dimensional model. In step 701, the facility receives a search request that specifies dimension values for one or more selected dimensions. Rather than specifying a value for each such dimension, in some embodiments, the search request may specify other tests with respect to a dimension, such as whether the dimension value is non-null, whether it falls within a particular range of dimension values or within a list of dimension values, whether it matches a pattern, or whether it satisfies an arbitrary programmatic or mathematical function.

[0043] Figure 8 is a diagram showing a sample search request. The search request 810 contains dimension values 822-824 provided by a user for the company_type, company_location, and company_category dimensions, respectively. This search request is for the company root pages of service provider companies in the satellite communications area that are located in California.

[0044] Figure 9 is a diagram showing a representation of the search request shown in Figure 8 expressed in a search request star notation. The search request star shown is comprised of nodes 910, 920, 930, 940, and 950. The search request star shown specifies the value "service provider" for the company_type dimension in node 930, specifies the value "California" for the company_location dimension in node 940, and specifies the value "satellite communications" for the company_category dimension in node 950.

[0045] Figure 10 is a diagram showing a second sample search request for the same dimensional model. The diagram shows a search request for the company root pages of companies whose names contain the word "if". Figure 11 is a

diagram showing the same search request expressed in search request star notation. Search requests such as this one may be executed using a variety of well-known searching techniques. These include regular expression matching techniques that may be coded and called from within a database server, such as an Oracle database server, as well as techniques using an SQL-like predicate. Those skilled in the art will recognize that additional techniques may also be employed.

[0046] Returning to Figure 7, the facility repeats steps 702-704 for each dimension selected in the search request. In step 703, the facility subsets the dimension table for the selected dimension down to the rows matching the dimension value specified in the search request for the selected dimension or satisfying the tests specified in the search request for the selected dimension. In other words, the facility selects the rows of the dimension table for the selected dimension whose dimension values match the specified dimension value. In step 704, if additional selected dimensions remain to be processed, the facility continues in step 702 to process the next selected dimension. When all of the selected dimensions have been processed, the facility continues in step 705. In step 705, the facility joins the dimension tables for the selected dimensions, as subsetted in step 703, to the fact table of the dimensional model. The result table produced by this join operation contains a row for each modeled document that satisfies the search request. Each row of the result table preferably contains a URL or other reference to the document satisfying the search, and, optionally, the values of attributes of each document, including attributes corresponding to dimensions not selected in the search request. For example, when each document is a book, the result table may include the current rank of the book on the New York Times bestseller list. In some embodiments, the dimension processing represented by steps 702-704 and the fact processing represented in step 705 are parallelized and proceed concurrently (not shown). In step 706, the facility displays a search result to the user based on the table produced by the

join operation. After step 706, the facility continues in step 701 to process the next search request.

[0047] As will be appreciated by those skilled in the art, the facility may be employed to generate and respond to search requests for dimensional models for a wide variety of document groups and the attributes of the constituent documents. Figures 12-14 show a few examples for other groups of documents.

[0048] Figure 12 is a diagram showing a search request, expressed in search request star notation, for yellow pages entries, *i.e.*, entries in a business telephone directory. As can be seen from nodes 1220, 1230, 1240, and 1250, the model for yellow pages entries models the entries on the following attributes: vendor_name, vendor_location, product_type, and payment_types. As can be seen from nodes 1220, 1230, and 1240, the query shown is for vendors whose names contain the word "Roma", whose locations contain the word "Colorado", and whose product types include the word "pizza".

[0049] Figure 13 is a diagram showing a search request, expressed in search request star notation, for descriptions of products sold by a computer superstore. As can be seen from nodes 1320, 1330, 1340, 1350, and 1360, the model for computer superstore product descriptions models the descriptions on the following attributes: vendor, product_name, product_type, operating_system, and price. As can be seen from nodes 1320 and 1340, the query shown is for computer superstore product descriptions for microphone products from Dragon Systems.

[0050] Figure 14 is a diagram showing a search request, expressed in search request star notation, for a library resource, such as a book or an audio or video recording. As can be seen from nodes 1420, 1430, 1440, 1450, and 1460, the model for library resources models library resources on the following attributes: title, author, publisher, subject, and publication_date. As can be seen from nodes 1430 and 1460, the query shown is for library resources published in 1998 by an author named Henley.

[0051] It will be understood by those skilled in the art that the above-described facility could be adapted or extended in various ways. For example, the facility may be used to model and select documents, other data artifacts, or other resources of virtually any type, including programmatic objects, such as COM or CORBA components or Java applets. While the foregoing description makes reference to preferred embodiments, the scope of the invention is defined solely by the claims that follow and the elements recited therein.